Interspeech 2018
2-6 September 2018, Hyderabad

Predicting Categorical Emotions by Jointly Learning Primary and Secondary Emotions Through Multitask Learning

Reza Lotfian and Carlos Busso

Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
reza.lotfian@utdallas.edu, busso@utdallas.edu

Abstract

Detection of human emotions is an essential part of affect-aware *human-computer interaction* (HCI). In daily conversations, the preferred way of describing affects is by using categorical emotion labels (e.g., sad, anger, surprise). In categorical emotion classification, multiple descriptors (with different degrees of relevance) can be assigned to a sample. Perceptual evaluations have relied on primary and secondary emotions to capture the ambiguous nature of spontaneous recordings. Primary emotion is the most relevant category felt by the evaluator. Secondary emotions capture other emotional cues also conveyed in the stimulus. In most cases, the labels collected from the secondary emotions are discarded, since assigning a single class label to a sample is preferred from an application perspective. In this work, we take advantage of both types of annotations to improve the performance of emotion classification. We collect the labels from all the annotations available for a sample and generate primary and secondary emotion labels. A classifier is then trained using multitask learning with both primary and secondary emotions. We experimentally show that considering secondary emotion labels during the learning process leads to relative improvements of 7.9% in F1-score for an 8-class emotion classification task.

Index Terms: Emotion recognition, categorical emotions, subjective evaluations, secondary emotions, multitask learning.

1. Introduction

The perception of emotion is an essential communication skill. Describing emotions with discrete categories such as happiness or anger is useful in human-human and human-computer interactions. For example, an affect-aware game that can measure if the player is happy with the difficulty of the game can make the game more enjoyable [1,2]. An interactive system that can communicate emotions can play a key role in teaching social interactions to children with autism spectrum disorders [3]. The key challenge is that emotions in spontaneous human interactions are ambiguous, so the boundary between categorical emotions is not clear [4].

In contrast to acted renditions, spontaneous emotions are elicited as a result of the interaction. These recordings are commonly annotated with perceptual evaluations, where several evaluators are asked to assign an emotional class to each sample. Depending on the evaluator's perspective, multiple answers may be appropriate, especially if many related emotional categories are available (e.g., surprise, anger, fear, disgust) [5]. Since the boundaries between emotional classes during human interaction are ambiguous, an individual listener may need more than one category to accurately describe their perception of the speaker's emotional state [6]. The main focus of previous studies was to address the ambiguity of emotions by taking into consideration conflicting labels reported by different annotators. While these approaches have shown to be successful in increasing the performance of emotion classifiers [7–9], they overlooked the case where multiple emotions are reported by a single evaluator. Listeners are often provided with the option to select extra classes that they find relevant during the subjective evaluations.

Devillers et al. [10] propose to label emotion with major and minor emotions. Major emotions correspond to conventional classes used to describe emotions (fear, anger, sadness, hurt, surprise, positive and neutral). Minor emotions were added to capture emotional traits not included in the major emotions. The list of the emotional classes was extended from 7 to 21 classes. We have also used a similar scheme for the annotation of the MSP-IMPROV [11] and MSP-Podcast [12] databases, where our evaluators are asked to provide a primary emotion (the most relevant class) and secondary emotion (all emotional classes perceived in the stimulus). Despite the fact that in these cases the information about secondary emotions is available, the problem is commonly formulated to identify the most relevant emotion (i.e., primary emotion). Therefore, the secondary emotion labels are ignored in training classifiers. We hypothesize that this information is useful and can be used to improve emotion classification.

This paper proposes to combine primary and secondary emotion labels using *multitask learning* (MTL) framework. Our approach treats primary and secondary labels as different tasks. Inspired by the emotion perception evaluation used to annotate a corpus, the primary classification task is to find the most relevant category. The secondary task is to define all the labels that are relevant. This approach is implemented with a MTL framework implemented with *deep neural network* (DNN) to jointly predict labels of primary and secondary emotions. The cost function for the primary emotion uses *cross-entropy* (CE), as this task is a conventional classification problem. The cost function for the secondary emotion uses the *Kullback-Leibler divergence* (KLD), since multiple classes are allowed (e.g., we use soft-labels instead of one-hot vectors). For the secondary task, we propose a sigmoid function that adjusts its parameters during training to identify emotional traits also included in the stimulus. We set the parameters of the MTL framework to balance the tradeoff between the primary and secondary task cost functions, maximizing the performance on the primary emotion task. We experimentally show that that considering secondary emotions in the model leads to significant improvements on the classification performance of the primary class.

This work was funded by NSF CAREER award IIS-1453781.

10.21437/Interspeech.2018-2464

2. Related Work

Previous work has demonstrated that *multitask learning* (MTL) technique can be used to jointly model both gender and emotion [13, 14], resulting in consistent improvement in the classification performance over gender-agnostic models. Previous work has also showed that emotion recognition systems could be improved by incorporating speaker's identity as a feature, along with other emotion-related features [15]. Zhang et al. [16] explored MTL frameworks for leveraging data from different domains (speech and song) and gender in emotion recognition systems. Xia et al. [17] treated dimensional and categorical labels as two different tasks. Recent studies conducted on dimensional emotion recognition have shown that jointly learning multiple attributes (e.g., arousal, valence, dominance) leads to overall improvement in the performance of regression [18] and classification of discrete classes [19].

MTL is best suited to situations where it is possible to train with all data from scratch. Its main advantage is to build a feature representation that is common between tasks, creating powerful regularization constraints that generalize better to new domains. While studies have used MTL for speech emotion recognition [15–22], the contribution of this study is the use of the annotations of secondary emotion, which are commonly discarded. The MTL framework considers the recognition of the single most relevant label as the primary task. The secondary task is to determine all the relevant labels included in the stimuli. This formulation regularizes the network, leading to significant improvements in the classification of the primary task. The proposed approach is extensively evaluated using a large naturalistic database, where annotations from primary and secondary tasks are available from multiple annotators.

3. Resources

3.1. The MSP-Podcast Database

This study uses the version 1.1 of the MSP-Podcast emotional speech corpus collected at the University of Texas at Dallas [12]. This database includes an extensive set of speech segments from podcast recordings available in audio sharing websites. The topics of the podcasts in this database include talk shows and discussions about movies, politics and sports, covering a diverse range of emotions. Then, speaker diarization tools are used to split podcasts into speech turns. Among the speech segments, only the ones with a single speaker, and without noise or background music are automatically selected. We also discard segments with phone quality audio. The duration of the segments are restricted to be between 2.75s and 11s. Since most of the speech turns are emotionally neutral, we rely on emotion retrieval systems to identify emotional speech. Following the ideas described in Mariooryad et al. [23], we use different machine-learning formulations to retrieve segments from the pool of available segments conveying target emotion behaviors. In the final step before the subjective evaluations, we manually screen all the retrieved samples to find the ones that do not satisfy the required criteria. The speech segments are annotated with emotional labels using an improved version of the crowd-sourcing method proposed by Burmania et al. [24] (see details on Lotfian and Busso [12]). The data collection process is an ongoing effort, where the current study uses 22,630 speech segments (39 hrs, 12 min). We have manually annotated the speaker identity of 284 speakers (19,007 segments). We use segments from 50 speakers as our test set (7,181 segments), and data from 20 speakers as our development set (2,614 segments). All the remaining samples are dedicated to the train set (12,835 segments). This set includes data from the remainder 214 speakers, plus the speaking turns that have not been labeled with speaker information. This partition attempts to create speaker independent datasets for the train, test and development sets.

During the annotation of categorical emotions, the raters are asked to choose the primary emotion from anger, sadness, happiness, surprised, fear, disgust, contempt, and neutral. They can also choose other if none of the previous labels are suitable. The raters can only select one class. Next, the annotators are asked to evaluate secondary emotions, where they are free to select all the emotions that they believe are relevant to the speech sample, including the classes selected for the primary emotions. The options for the secondary emotions are extended by adding the following emotions: amused, frustrated, depressed, concerned, disappointed, excited, confused, and annoyed. Each speech segment is annotated by at least five annotators. While the corpus also has evaluations for attribute-based descriptors, this study only uses primary and secondary categorical emotions.

3.2. Acoustic Features

We use the *Geneva minimalistic acoustic parameter set* (eGeMAPS) feature set [25] designed for paralinguistic tasks. The eGeMAPS contains a total of 88 *high level descriptors* (HLDs) including frequency, energy, spectral, cepstral, and dynamic information. This set was carefully selected based on their performance in previous paralinguistic problems, and theoretical significance. The final feature vector for each utterance is obtained by applying different statistics (see details on Eyben et al. [25]). Because of the minimalistic size of the set, we use all 88 features without applying feature selection. This approach facilitates the reproduction of the results by other groups. The features are extracted with OpenSMILE [26]. We perform z-normalization on each feature based on the parameters estimated over the train set.

4. Multitask Learning Architecture

The proposed framework to leverage annotations of secondary emotions relies on multitask learning. The key idea is to simultaneously solve primary and secondary emotions, creating a robust feature representation for this task. We formulate the prediction of emotional categories as multi-class classification problem that is solved with DNNs. The task is an 8-class problem (anger, sadness, happiness, surprised, fear, disgust, contempt, and neutral). The network takes the acoustic features (Sec 3.2) as an input vector and maps it into an output vector while each element represents the relevance of the input features to the corresponding emotional category. In our formulation, the primary objective is to correctly predict the probabilities of the most relevant class (i.e., primary emotion). The secondary or auxiliary task is to predict all the emotional categories that are relevant to the affective content of the input speech. This problem is formulated as a detection task per emotion class (i.e., multiple classes are possible). The primary and auxiliary tasks are jointly learned in the network. Therefore, we optimize the network to solve for both problems.

The ground truth for secondary emotion labels are generated by combining the secondary emotions selected by all the annotators. First, we remove the secondary labels that are not among the list of primary emotion categories. Therefore, the number of categories for primary and secondary classes are the

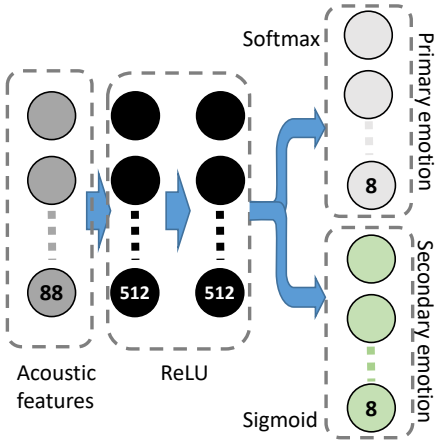

Figure 1: *The proposed framework to jointly recognize primary and secondary emotions. The framework has two fully connected layers, each of them with 512 nodes.*

same. Then, we find the average number of secondary emotions selected for the sentence and find its closest integer (k). Next, we select the k most frequent classes selected by the annotators. These emotional classes are set to one. The rest of the classes are set to zero. Notice that the primary emotion is always included as one of the secondary classes. The size of the generated label vector is equal to the number primary emotion classes. Unlike the labels for the primary emotion, where only one class is allowed, the ground truth for the secondary emotions has k classes.

Figure 1 shows the architecture of the proposed MTL network. The first two hidden layers are shared between the primary and secondary tasks. The layers are then connected to the output layers for the primary and secondary tasks. The primary emotion output is compared to the ground truth using a softmax layer with cross-entropy as the loss function. We generate the secondary emotion outputs using a sigmoid layer. For training, we use the *Kullback- Leibler divergence* (KLD) between the ground truth vector and the predicted vector as the loss function for secondary emotions. This approach generates two different losses for primary and secondary emotions. We combine the primary and secondary emotion losses to generate the overall optimization loss by calculating the weighted sum as shown in equation 1:

$$L_{ov} = (1 - \alpha) \times L_{primary} + \alpha \times L_{secondary} \quad (1)$$

The value of the parameter α is set to maximize the classification performance of the primary task over the development set. We compute the precision and recall rates for all the eight classes. Then, we estimate the mean precision ($\bar{P}$) and recall ($\bar{R}$) rates. These values are used to estimate the F1-score using Equation 2.

$$F1 - score = \frac{2\bar{P}\bar{R}}{\bar{P} + \bar{R}} \quad (2)$$

5. Experimental Evaluations

We refer to our method as *MTL (PE + SE)*, since it considers *primary emotions* (PE) and *secondary emotions* (SE). The ground truth labels for the test set are assigned by finding the majority vote from the primary emotions assigned by the annotators. The consensus label is compared with the one-hot label generated from the output of the primary task model by setting the node with the highest value to 1 and rest of the nodes to 0.

5.1. Baseline Frameworks

In addition to our proposed method, we implement two other machine-learning approaches as baselines. The first method is training using the majority vote obtained from the primary emotion labels. We call this method *hard label primary emotion* (Hard label (PE)). The second baseline uses soft-labels derived from the primary labels followed the method described by Fayek et al. [8]. This method creates a label vector for the emotional classes considering the distribution of the emotions assigned to each sample. The output layer is implemented with softmax and cross-entropy for the cost function. We refer to this baseline as *soft-label (PE)*. We consistently implement these baselines with two hidden layers and 512 nodes in each layer.

5.2. Results

The parameter α of the MTL network is optimized over the development set to maximize the performance of the primary task. Figure 2 shows the F1-score of the classification on the development set with different values for α. The best value for α is $\alpha = 0.56$, which assigns 44% weight to the primary task and 56% to the secondary task. We set α to this value in the evaluation on the test set.

We compare our proposed method to the baselines. Table 1 shows unweighted average precision, recall and F1-score of the classification using the proposed MTL and the two *single-task learning* (STL) baselines. We also include the performance of human annotations to have a reference for the difficulty of the task. The human performance is obtained by excluding one of the annotators at random for each sample. We find the consensus label from the remaining annotators. Then, we compare the excluded label to the consensus label obtained from the rest of the evaluators. The relatively low F1-score of 38.6% shows the difficulty of this classification task (chance performance is 12.5%). Notice that the MSP-Podcast corpus is a collection of spontaneous emotional interactions, as opposed to prototypical behaviors. The classes are not balanced which make the classification problem more challenging. While the classifiers only rely on acoustic features, human annotators use semantic information which add discriminative information to assess the emotional content.

Considering secondary emotions using MTL leads to improvements in the F1-score over the two other STL networks. We estimate the statistical significance of the improvements by using a one-tailed z-test on the difference in population proportions, asserting significance if p-value ≤ 0.05. The results reveal that the differences are statistically significant, indicating that the proposed MTL outperforms the baselines. The conventional approach is to train the DNN with hard labels (*Hard label (PE)*). Our MTL approach using primary and secondary emotions achieves an absolute improvement in F1-score of 2.3%, which represents a relative gain of 9.6%.

We also compare the performance of the classifiers in terms of the loss function for the primary emotion classification task on the test set, using the cross-entropy measure. Table 2 shows the results for the primary emotion task. In spite of considering an auxiliary task in the training of the network, the loss is smaller than the baselines. This result suggests that using infor-

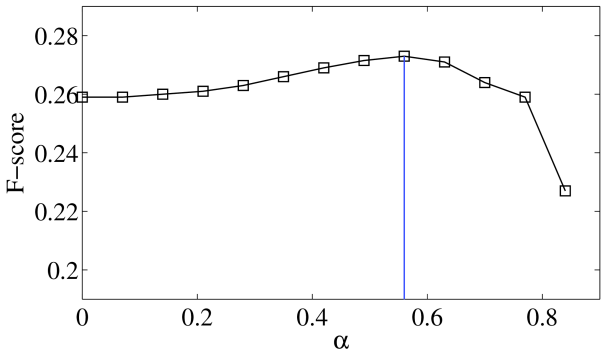

Figure 2: *F1-score performance as a function of α (Eq. 1).*

Table 1: *The unweighted precision, recall and F1-score of the primary emotion classification using different machine-learning frameworks. Double asterisks (**) shows that the approach outperforms other alternative methods. Single asterisk (*) shows the approach outperforms the Hard-label PE baseline. We assert significance if p-value$\leq$0.05.*

	Precision [%]	Recall [%]	F1-score [%]
Hard-label PE	23.1	24.9	24.0
Soft-label PE	24.9	25.8	25.3*
MTL (PE+SE)	26.4	26.1	26.3**
Human performance	40.8	37.2	38.9

mation from the secondary emotion leads to better generalization.

We also report the performance of the secondary task. We highlight that the MTL framework is trained to maximize the classification performance of the primary task. Therefore, the performance of the secondary task plays a lesser role during training. Since multiple secondary emotions are possible, we evaluate this problem as a binary classification task per emotional class, determining whether the speech in the test set is relevant to each class. We find the accuracy for each emotional class, which are aggregated by estimating the unweighted accuracy across all classes. In this evaluation, we compare the proposed MTL framework with a STL framework that recognizes secondary emotions. We refer to this baseline as *Hard-label SE*. Note that the *Hard-label PE* and *Soft-label PE* baselines predict only the primary emotion, so we cannot use them as baselines in this evaluation. Table 3 shows the average accuracy for the two methods. The results show that the proposed MTL framework outperforms the STL baseline in detecting secondary emotions task by 5.1%. The difference is statistically significant (p-value ≤ 0.05). This baseline method is only trained to maximize the performance in detecting secondary emotions, which is an unfair advantage over the MTL, which is optimized to maximize the classification performance of the primary emotion. In spite of this disadvantage, the proposed MTL framework still provides better results for the secondary task. This result demonstrates that the shared feature representation learned by the model is discriminative for both tasks, leading to improved classification performance over STL methods.

Table 2: *The average cross-entropy loss on the test set for the primary classification task for different networks.*

	Loss
Hard-label PE	1.391
Soft-label PE	1.350
MTL (PE+SE)	1.339

Table 3: *The unweighted accuracy in detecting secondary emotion on the test set.*

	Accuracy
Hard-label SE	61.7%
MTL (PE+SE)	66.8%

6. Conclusions

Categorical emotions are a common way of describing emotions in daily conversations. The challenging aspect of using categorical emotions in machine-learning is the ambiguity between classes. When individuals are asked to describe their perceived emotion, they often use multiple categorical descriptors. From an application perspective, however, detecting the single most relevant class is usually preferred. Therefore, classifiers are often trained with the information from the most relevant class, discarding other categories perceived by the evaluators. This study proposed a multitask learning solution to improve the performance of recognizing the primary emotional class by leveraging extra information provided in the evaluations about secondary emotions. We experimentally show that multitask learning increases the classification performance of the primary emotion, and that the improvements are statistically significant.

While the use of MTL is not new in speech emotion recognition, the use of secondary emotion labels to improve the classification performance of a primary emotion is a novel concept, which is systematically explored in this study. Considering this extra information, which is often discarded by other studies, leads to improvements over conventional approaches.

As a direction for future research, it is possible to include attribute-based emotions (e.g., arousal, valence and dominance) as extra auxiliary tasks in addition to the secondary emotions. Another possible research direction is to investigate the optimum criteria to accept an emotion as secondary, when multiple annotations exist. We will study the cut-off threshold for considering relevant categories.

7. References

[1] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in *Image Processing & Communications Challenges 6*, R.S. Choraś, Ed., vol. 313, pp. 227–236. Springer International Publishing, Bydgoszcz, Poland, September 2014.

[2] M. Obaid, C. Han, and M. Billinghurst, ""Feed the fish": an affect-aware game," in *Australasian Conference on Interactive Entertainment*, Brisbane, Australia, December 2008.

[3] B. Abirached, Y. Zhang, and J.-H. Park, "Understanding user needs for serious games for teaching children with autism spectrum disorders emotions," in *World Conference on Educational Media and Technology (EdMedia 2012)*, Denver, CO, USA, June 2012, pp. 1054–1063.

[4] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S.S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing*

and Intelligent Interaction (ACII 2009), Amsterdam, The Netherlands, September 2009, pp. 1–8.

[5] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.

[6] K.R. Scherer, "Appraisal theory," in *Handbook of cognition and emotion*, T. Dalgleish and J.M. Power, Eds., pp. 637–663. John Wiley & Sons Ltd, New York, NY, USA, March 1999.

[7] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 490–494.

[8] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *International Joint Conference on Neural Networks (IJCNN 2016)*, Vancouver, BC, Canada, July 2016, pp. 566–570.

[9] K. Audhkhasi and S. S. Narayanan, "Emotion classification from speech using evaluator reliability-weighted combination of ranked lists," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4956–4959.

[10] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[11] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[12] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2018.

[13] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *European Signal Processing Conference (EUSIPCO 2004)*, Vienna, Austria, September 2004, pp. 341–34.

[14] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *International conference on Language Resources and Evaluation (LREC 2006)*, Genoa,Italy, May 2006, pp. 1123–1126.

[15] M. Sidorov, S. Ultes, and A. Schmitt, "Comparison of gender-and speaker-adaptive emotion recognition," in *International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, May 2014, pp. 3476–3480.

[16] B. Zhang, E. Mower Provost, and G. Essi, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5805–5809.

[17] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, January-March 2017.

[18] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[19] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4490–4494.

[20] J. Chang and S Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2746–2750.

[21] J. Kim, G. Englebienne, K.P. Truong, and V. Evers, "Towards speech emotion recognition "in the Wild" using aggregated corpora and deep multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1113–1117.

[22] D. Le, Z. Aldeneh, and E. Mower Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1108–1112.

[23] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[24] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[25] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.